African Journal of Biotechnology Vol. 10(57), pp. 12147-12155, 28 September, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.1761
ISSN 1684–5315 © 2011 Academic Journals

Full Length Research Paper

De novo characterization of *Lycoris sprengeri* transcriptome using Illumina GA II

Le Chang, Jingjue Chen, Yumian Xiao and Yiping Xia*

Department of Horticulture, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China.

Accepted 29 August, 2011

The Illumina GA II high-throughput sequencing technology was used to establish a transcriptome library from the bulbs of *Lycoris sprengeri*. A total of 24 million short read with an average length of 90 nt were generated, which produced 98,150 unigenes with a mean length of 385 nt. A total of 45.9% of these unigenes (45,052) were annotated using BLAST (E-value ≤1.0E^{-5}) against the nr, SwissProt, Kyoto encyclopedia of genes and genomes (KEGG) and Clusters of orthologous groups (COG) databases. The unigenes were annotated with gene ontology (GO) and COG. Of these, 2,173 unigenes were found to be related to carbohydrate metabolism, and 544 were involved in starch and sucrose metabolism.

Key words: *Lycoris sprengeri,* Illumina GA II, transcriptome, carbohydrate metabolism.

INTRODUCTION

The plants of the genus *Lycoris*, also known as 'Chinese tulips', originated and are mainly distributed in the east of China. They bear showy and elegant umbels with narrow reflexed petals, possess a special flowering habit, and have great medicinal value (Bryan, 1992). Moreover, they grow in harsh conditions and can be easily used in landscaping. In contrast to the well known characteristics of bulbous vegetative and reproductive growth, the developmental mechanisms of *Lycoris* bulbs are unclear (Wang et al., 2009). The only way to elucidate the mechanisms of *Lycoris* development is to describe the genes responsible for different developmental stages, but there are few of such studies on the *Lycoris* genome or transcriptome. Only cDNA libraries of *Lycoris aurea* leaves and flower buds, and petals of *Lycoris longituba* have been constructed (Cui, 2004; He et al., 2006). Their expressed sequence tags (ESTs) were also sequenced and analyzed by bioinformatics. These studies lay a solid foundation for genetic analysis. In the NCBI database, however, only trace sequence information of *Lycoris* can be found. Information on nuclear and mitochondrial coding regions are included, but these data are insufficient to study gene function.

We sequenced the transcriptome of *Lycoris sprengeri* using the next generation of high-throughtput paired-end RNA sequencing (RNA-seq) technology, Illumina Genome Analyzer (GA) II, which lacks genome information (Wang et al., 2009). One of the most important features of Illumina GA II is the transcript assembly independent of genome (Ansorge, 2009). The previous reports on *de novo* assembly could refer to researches in whitefly, brown planthopper and marine fish (Wang et al., 2010; Xue et al., 2010; Xiang et al., 2010), whose genomic sequence resources were unknown. Millions of reads were generated from a single sample per run, many more than the thousands of ESTs produced by EST sequencing (Cui, 2004; Delseny et al., 2010). In addition, this sequence analysis covered almost all of the mRNA in a single cell (Cirulli et al., 2010). Sequences with extremely low abundance could also be detected (Rosenkranz et al., 2008; 't Hoen et al., 2008). RNA-seq can also reveal gene expression patterns in certain organs or during different developmental stages (Schuster, 2008; Ansorge, 2009). For plants, RNA-Seq based on Illumina GA II was reported in *Oryza sativa* (Zhang et al., 2010), *Arabidopsis thaliana* (Filichkin et al., 2010) and *Artemisia annua* (Graham et al., 2010) but not for flowering bulbs.

In this study, the Illumina GA II was used to sequence the transcriptome of *L. sprengeri*. Then, bioinformatics technology was used to perform a *de novo* assembly and annotation without prior genome information. This transcriptome database helped to reveal much about the functional genomics of *L. sprengeri*, and was then used

*Corresponding author. E-mail: ypxia@zju.edu.cn.

Table 1. Illumina sequencing and assembly quality of *L. sprengeri* transcriptome.

Parameter	Value
Total number of clean reads	24,021,994
Total nucleotides (nt)	2,161,979,460
Average read length (nt)	90
Total number of contigs	875,126
Mean length of contigs (nt)	120
Total number of scaffolds	185,571
Mean length of scaffolds (nt)	268
Total number of unigenes	98,150
Mean length of unigenes (nt)	385

to predict the functional classification of many unigenes using GO, COG classification and KEGG pathway analysis. These results lay the foundation for understanding the relation between gene expression patterns and plant develop-ment, physiology and structure, and will be helpful for the molecular breeding of *L. sprengeri*. Furthermore, we focused on the sequences that are related to carbohy-drate metabolism in the aim of exploring the relationship between changes in carbohydrate metabolism and bulb development.

MATERIALS AND METHODS

Sample preparation and RNA isolation

Four- to five-year-old mature bulbs of *L. sprengeri* of similar size were collected after flowering from the Zhoushan Islands (E 121°30'-123°25', N 29°32'-31°04') of Zhejiang province, China. The average circumference of the bulbs was 15.02 cm and the average weight was 69.70 g. The fresh samples were cleaned, cut into pieces, and used immediately to extract total RNA. The total RNA was isolated from the bulbs using the SV Total RNA Isolation System (Promega) according to the manufacturer's instructions.

cDNA library construction and sequencing

The total RNA was treated with RNase-free DNase I for 30 min at 37℃ (New England BioLabs) to remove residual DNA. The RNA Integrated Number (RIN) was 8.3 according to the Agilent 2100 check, so the total RNA was of sufficient quality. Then, cDNA library was prepared using the Illumina kit following the manufacturer's protocol (Zhang et al., 2010). Poly(A) mRNA was isolated from the total RNA using oligo (dT) magnetic bead adsorption. The mRNA was cut into short fragments (about 200 nt) by fragmentation buffer at 94℃ for 5 min. With these short fragments as templates, the first-strand cDNA was synthesized using random hexamer-primers, and the second-strand cDNA was synthesized using buffer, dNTPs, RNaseH and DNA polymerase I. Short fragments were purified with the QiaQuick PCR extraction kit and resolved with EB buffer for end reparation and addition of poly(A), and were then connected with sequencing adapters. After the agarose gel electrophoresis, suitable fragments were selected for PCR amplification as templates. Finally, the cDNAs were sequenced using Illumina HiSeq™ 2000. The size was approximately 200 bp and both ends were sequenced (Wang et al., 2010).

Reads purification and assembly

Sequencing-received raw image data was transformed by base calling into raw reads and stored in fastq format. Before data analysis, raw reads that had only 3' adaptor fragments, empty reads and reads with unknown sequences 'N' were removed. The clean reads were then randomly clipped into 31 bp oligomers. Transcriptome *de novo* assembly was performed using SOAPdenovo software (Li et al. 2010). Clean reads with a certain length of overlap were first joined into contigs. Then, the reads were mapped back to contigs, so that contigs from the same transcript as well as the distances between these contigs were detected by paired-end reads. Next, contigs were connected to form scaffolds. Paired-end reads were used again for gap filling of scaffolds to yield unigenes. In the final step, unigenes were aligned by Blastx to the protein databases like nr, Swiss-Prot, KEGG and COG (evalue < $1.0E^{-5}$). The best aligning results were used to decide the sequence direction of the unigenes, and the proteins with the highest sequence similarity with the given unigenes were retrieved (Xue et al., 2010).

Functional annotation of predicted proteins

With nr annotation, the Blast2GO program (Conesa et al. 2005) was used to get the G) annotation of unigenes, and WEGO software (Ye et al., 2006) was used for GO functional classification (molecular function, cellular component and biological process) for all unigenes. Unigenes were aligned by Blastall software to yield COGs and the KEGG database was used to predict the possible functional classification and the molecular pathways of unigenes.

RESULTS AND DISCUSSION

Illumina sequencing output statistics and reads assembly

A cDNA library from *L. sprengeri* bulbs was constructed and sequenced using the Illumina GA II system. After filtering the dirty raw reads (normally, the ratio of raw reads and clean reads is 0.6 - 0.65:1) and checking the quality, more than 24 million paired-end short reads of 90 nt in length were obtained from eight lanes of sequencing (Table 1). The total length of the reads was over 2.11 gigabases (Gb). These raw reads were randomly clipped into 31-mers, and combined into 875,123 contigs with a mean length of 120 nt. Then, using paired-end joining and gap-filling, the contigs were connected to yield 185,571 scaffolds (with mean length of 268 nt), in which about 26.7% had gaps. After gap filling of scaffolds, 98,150 unigenes (with mean length of 385 nt) were finally revealed (Table 1). The average coverage of unigenes was 84.26%, suggesting that the assembled unigenes covered most of the transcriptome sequences. The length distribution revealed that 82.03% of unigenes were between 100 and 500 nt (Figure 1C).

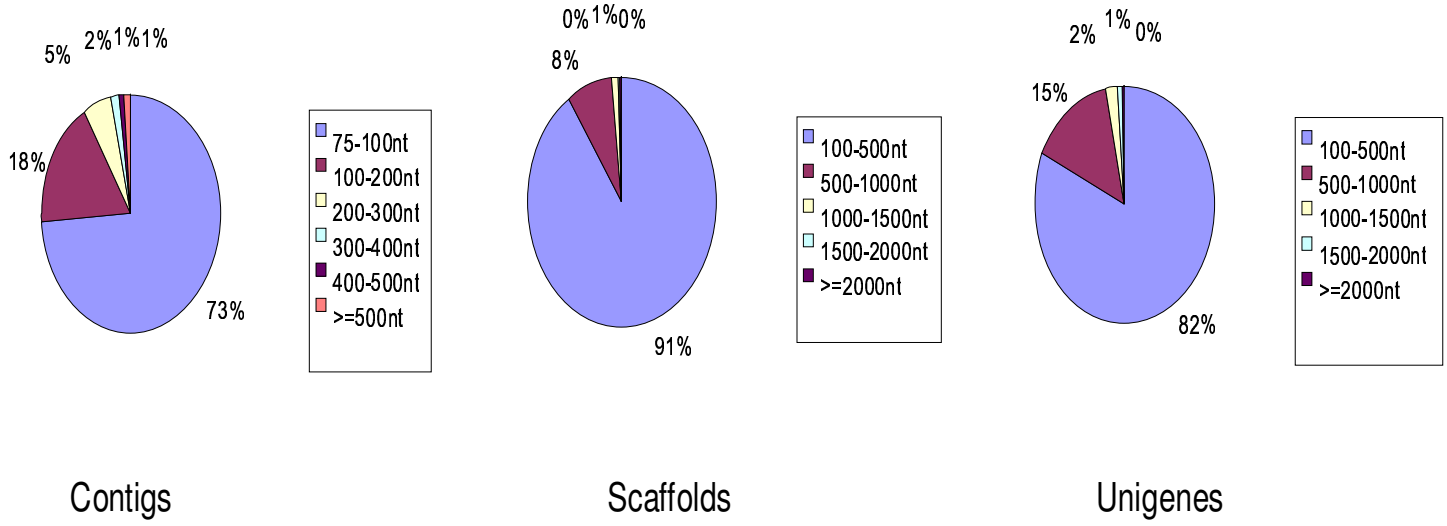

Figure 1. Length distribution of contigs, scaffolds and unigenes from *L. sprengeri* transcriptome sequencing and assembly.

Table 2. BLAST analysis results against important public databases.

Database	98,150 unigenes in library	
	Annotated (n)	Percentage (%)
nr	44,632	45.5
Swiss-Prot	26,637	27.1
KEGG	18,854	19.2
COG	11,155	11.4
Total	45,052	45.9

Unigene function annotation

Unigene sequences were aligned by Blastx to the non-redundant (nr) NCBI nucleotide database, as well as to Swiss-Prot, KEGG and COG (with the e-value of 10^{-5}). A total of 45,052 unigenes were matched, comprising 45.9% of all the unigenes in the transcriptome library (Table 2). Among these databases, most of the sequences were annotated against nr (44,632). Figure 2 indicates that the proportion of sequences with matches in the nr database was greater among the shorter assembled sequences, especially for lengths of 100 to 500 nt, with 66.7% match efficiency for annotated unigenes. The match efficiency gradually decreased with longer sequences. However, 54.1% of the unigenes still did not match any known genes in any database. These unigenes are probably too short, or are those in non-coding regions. The lack of referenced genome information is considered to be the main reason.

The characteristics of the homology search of Illumina sequences against the nr database were further analyzed. The E-value distribution of the top hits in the nr database indicated that 18% of the mapped sequences had strong homology (evalue <$1.0E^{-50}$) with other organisms (Figure 3). For the best-matched species distribution, 26.8% of the distinct sequences mapped to the sequences of *A. thaliana*, and 26.4% mapped to *O. sativa* (japonica cultivar-group), followed by *Zea mays* (8.2%) and *Vitis vinifera* (1.3%) while the remaining mapped sequences were distributed between $1.0E^{-5}$ and $1.0E^{-50}$, indicating little homology with other species in the database.

Highly expressed transcripts in the transcriptome library

The number of raw reads aligned to the unigenes represented the expression abundance. The top 10 most highly expressed transcripts are listed in Table 3. Except for the two that were not annotated in the databases, the other eight were annotated from sequences of *Z. mays*, *A. thaliana*, *Prunus persica*, *O. sativa* (Japonica cultivar-group), and *Hyacinthus orientalis*. The most highly expressed transcript was Unigene 47904, annotated as lipoxygenase (with 7810 raw reads), an important enzyme for fruit senescence, seed germination, plant development and the resistance response against stress (He et al., 2011). Unigene 47904 was followed by Unigene 19591, which was not found in any database. Unigene 51944, with 3344 raw reads, was annotated as GNS1/SUR4, an enzyme with important roles in glucose metabolism. The next most highly expressed was the DEAD-box RNA helicase-like protein (2133). The remaining transcripts were annotated as Os05g0498700 (transcribed locus), transaldolase, ABI3 interacting protein, binding and Os01g0767700 (Table 3).

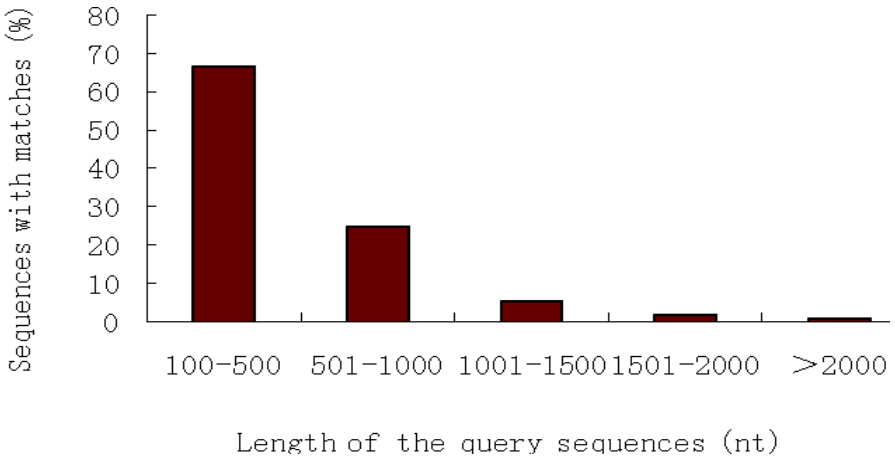

Figure 2. Effect of query sequence length on the percentage of sequences for which significant matches were found.

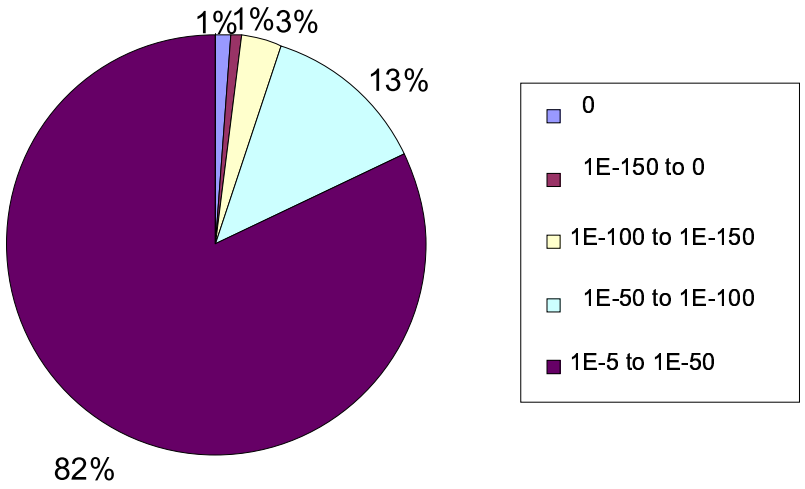

Figure 3. Characteristics of the homology search of illumina sequences against the nr database. E-value distribution of nr with a cut-off E-value of $1.0E^{-5}$.

Table 3. The ten most highly expressed transcripts in the *L. sprengeri* transcriptome library.

Gene ID	Raw reads	Nr-ID	BLAST annotation
Unigene47904	7810	gi|162459823|ref|NP_001105515.1|	lipoxygenase [*Zea mays*]
Unigene19591	7476	--	--
Unigene51944	3344	gi|15234538|ref|NP_195401.1|	GNS1/SUR4 membrane family protein [*Arabidopsis thaliana*]
Unigene24536	2133	gi|283049400|gb|ADB07168.1|	DEAD-box RNA helicase-like protein [*Prunus persica*]
Unigene66235	1615	gi|115464703|ref|NP_001055951.1|	Os05g0498700 [*Oryza sativa* (Japonica cultivar-group)]
Unigene1823	1424	gi|47027008|gb|AAT08720.1|	transaldolase [*Hyacinthus orientalis*]
Unigene67539	1269	gi|213959156|gb|ACJ54912.1|	ABI3 interacting protein [*Oryza sativa* Japonica Group]
Unigene64774	1260	--	--
Unigene1015	1231	gi|186523158|ref|NP_197072.3|	binding [*Arabidopsis thaliana*]
Unigene37961	1211	gi|115440165|ref|NP_001044362.1|	Os01g0767700 [*Oryza sativa* (japonica cultivar-group)]

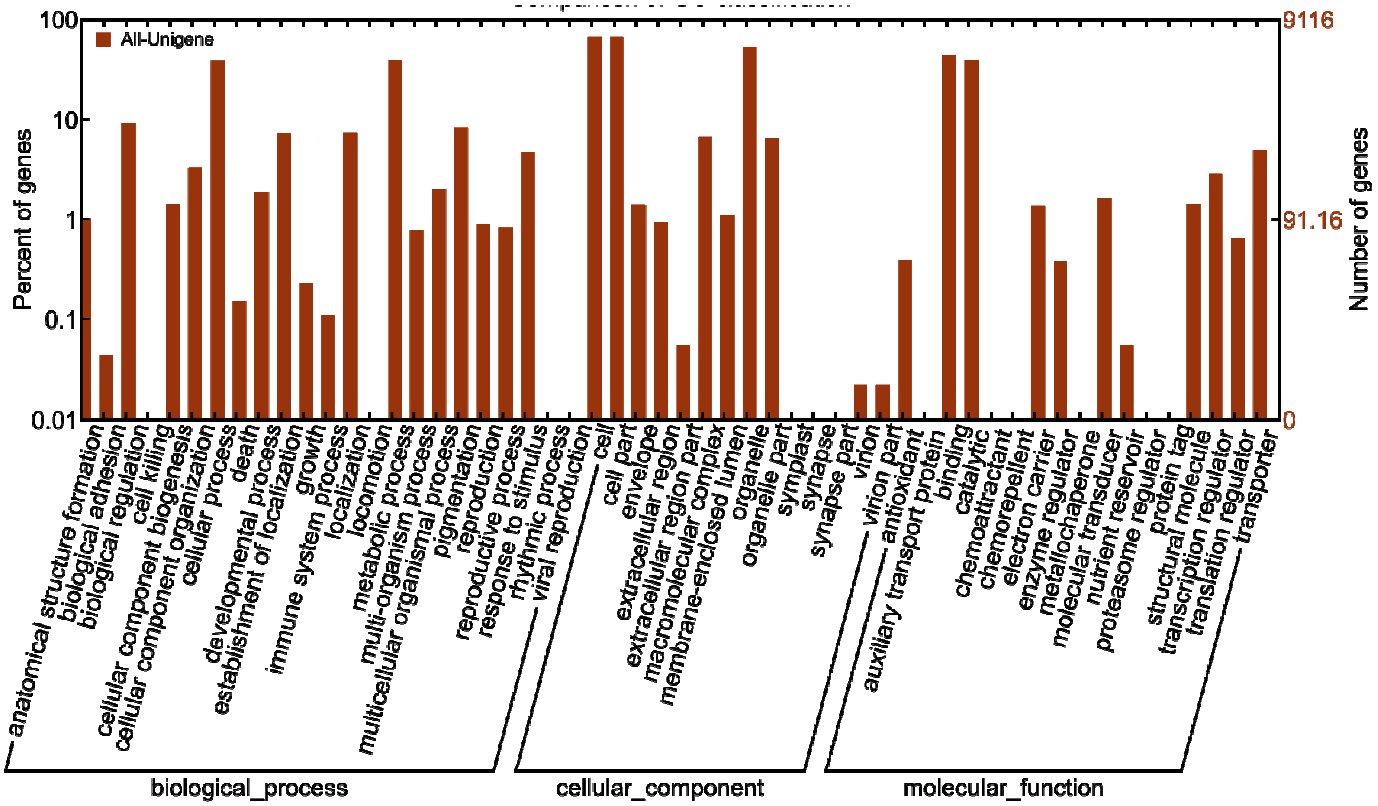

Figure 4. Comparison of the GO categories and subgroups of the unigenes. Unigenes were annotated into three categories: cellular components, molecular functions and biological processes.

GO and COG classification

According to the GO database, 39,129 sequences were classified into three categories: biological process (11,676), cellular component (18,605) and molecular functions (8,848). The sequences were divided into 52 functional groups. Some sequences might be involved in different regulating processes (Figure 4). In the category of biological process, 'metabolic process', 'cellular process' and 'biological regulation' comprised the largest proportion of sequences, accounting for 39.4, 39.2 and 9.1%, of the total, respectively. No sequence was categorized into the 'cell killing', 'locomotion', 'rhythmic process' or 'viral reproduction' subgroups. In the category of cellular components, 'cell part' and 'envelope' comprised 67.7% equally, and 'organelle' was 52.0%, and these three subgroups were dominant over the others. No gene in the 'symplast', 'synapse' and 'synapse part' groups were found. In the category molecular function, sequences with the functions of 'binding', 'catalytic' and 'transporter' comprised 44.1, 39.3 and 4.9% of the total. No genes corresponding to the functional subgroups 'auxiliary transport protein', 'chemoattractant', 'chemorepellent', 'metallochaperone', 'proteasome regulator', 'protein tag' were not found.

Unigenes were aligned to the COG database to predict and classify possible functions. In all, 19,071 out of 45,052 sequences had a COG functional classification. The sequences were classified into 25 COG categories (Figure 5). The single most common category was 'General function prediction' with 2,744 unigenes or 14.4% of all sequences. This was followed by 'transcription' (1771, 9.3%), 'replication, recombination and repair' (1771, 9.3%), and 'posttranslational modification, protein turnover, chaperones' (1587, 8.3%). The 'extracellular structures', 'nuclear structure' and 'cell motility' were the smallest COG categories.

The COG functional classification of *L. sprengeri* was then compared with *A. thaliana* and *O. sativa* according to the NCBI database (Figure 6). It can be seen from the figure that the unigenes with the predicted function of 'carbohydrate transport and metabolism' were always preponderant than other function such as 'amino acid transport and metabolism', 'nucleotide transport and metabolism', 'coenzyme transport and metabolism', 'lipid transport and metabolism' , with the exception of the top unigenes related to processing of germplasm.

Analysis of metabolic pathways

The 45,052 annotated unigenes were mapped to the

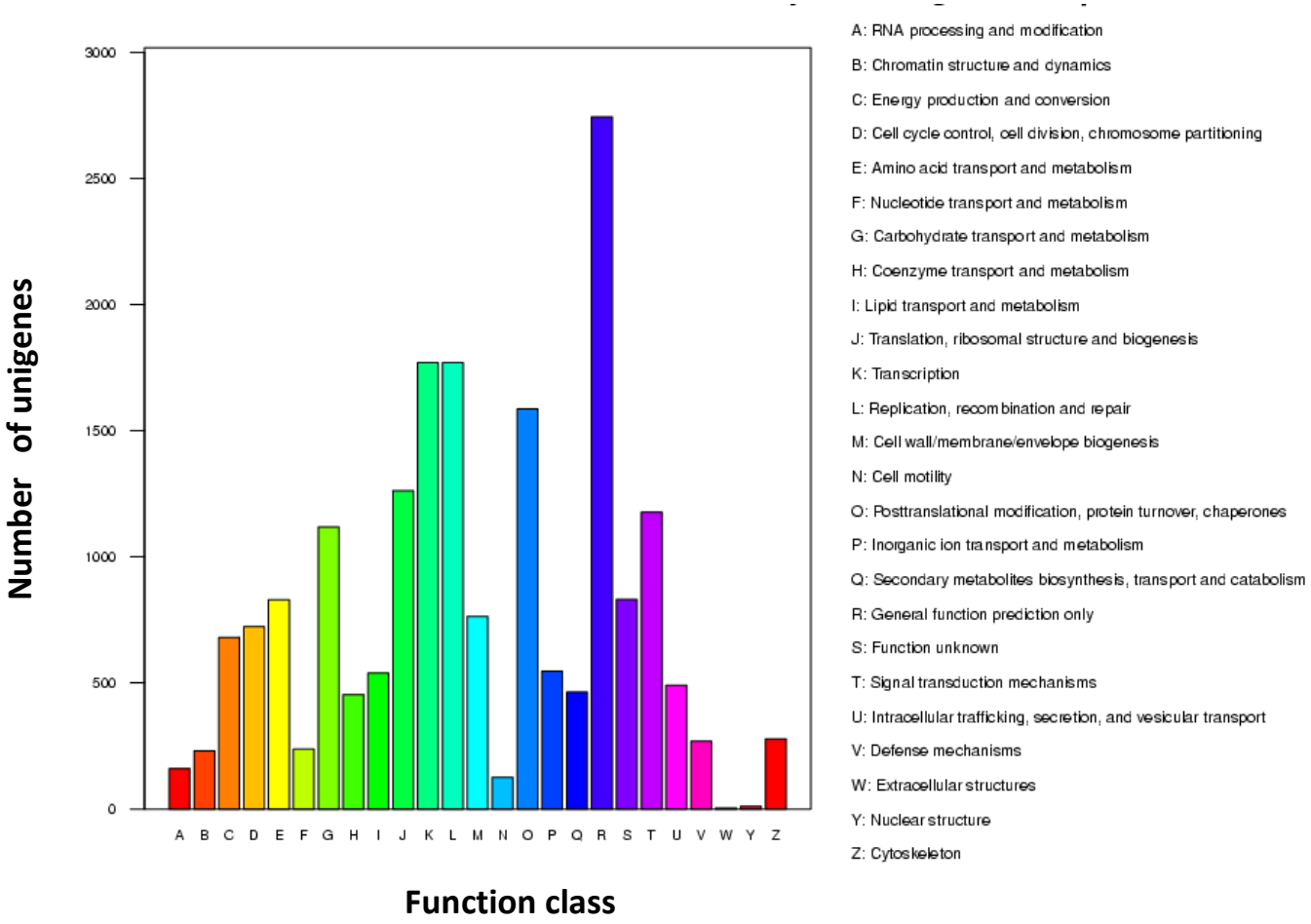

Figure 5. Number of *L. sprengeri*unigenes in the 25 COG functional classes.

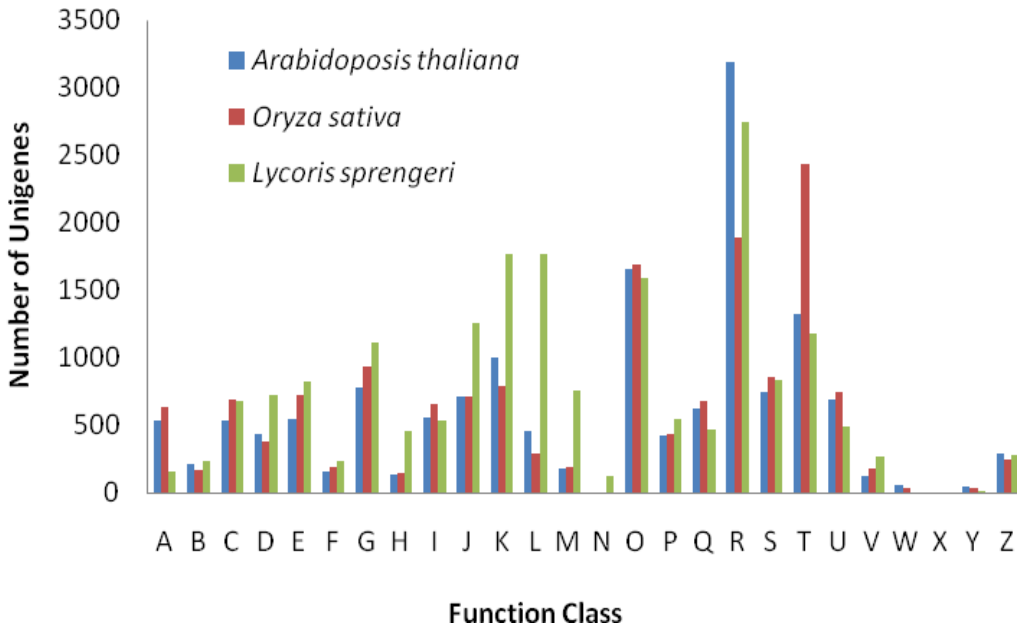

Figure 6. Comparison of COG functional classes-term among *A. thaliana, O. sativa* and *L. sprengeri*.

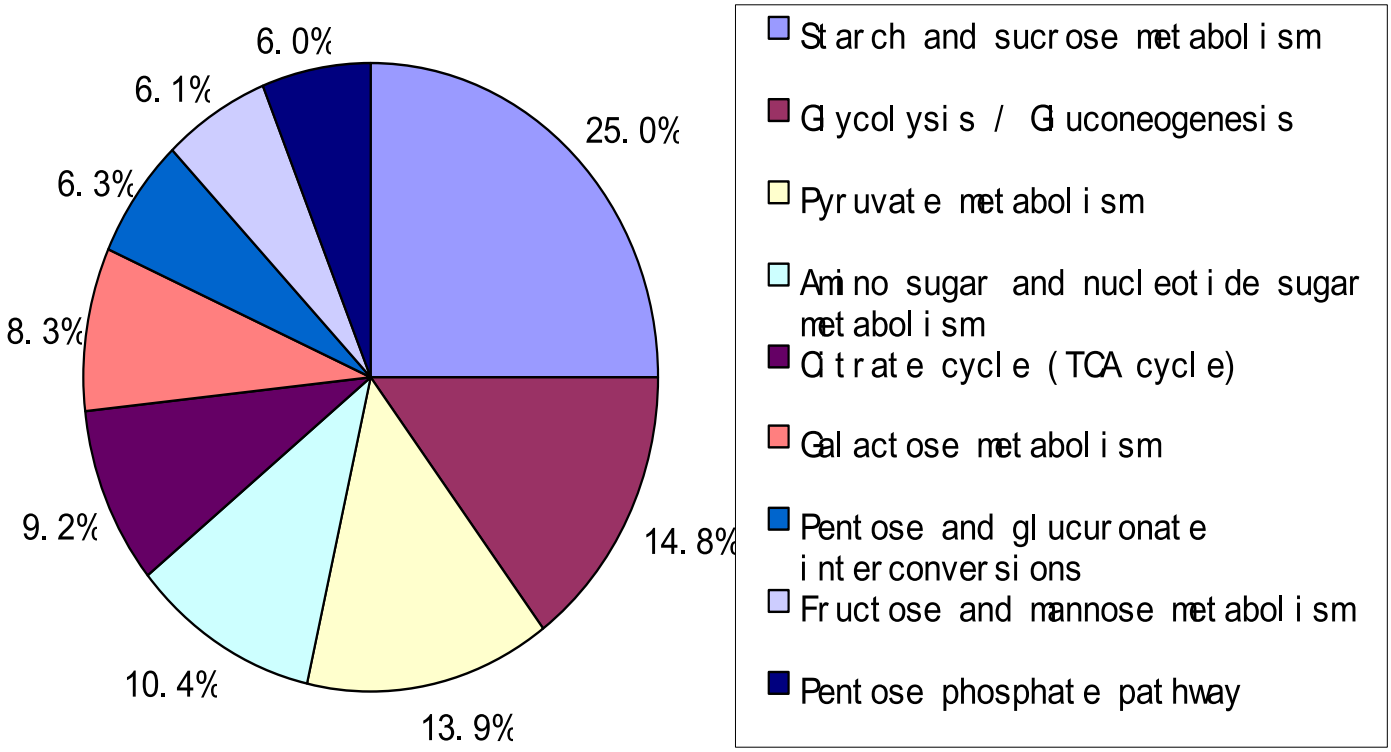

Figure 7. The proportion of sequences related to carbohydrate metabolism.

KEGG database to define the cellular pathways containing these unigenes. In total, 18,855 sequences were group into 125 pathways. The most dominant clusters were metabolic pathways (4570 members), plant-pathogen interaction (1254 members), spliceosome (1238 members), and biosynthesis of plant hormones (945 members). The bulbs of *L. sprengeri* were rich in starch, which is the main form of nutrient storage. The accumulating of starch and soluble sugar correlated with the beginning of bulb development (Sun et al., 2005). Thus, the study of carbohydrate metabolism and the developmental expression patterns of genes involved in carbohydrate metabolisms will reveal much about the developmental course of *L. sprengeri* bulbs. In total, 2,173 unigenes were associated with carbohydrate metabolisms according to KEGG. They could be subdivided further into genes involved in starch and sucrose metabolism, glycolysis/gluconeogenesis, pyruvate metabolism, amino sugar and nucleotide sugar metabolism, TCA cycle, galactose metabolism, pentose and glucuronate interconversions, fructose and mannose metabolism, and pentose phosphate pathway among others (Figure 7).

The detailed processes involved in the starch and sucrose metabolic pathways (the most common carbohydrate metabolic pathways) accounted for 25% of all these unigenes (Figure 7). According to the BLAST analysis results, 33 key enzymes (outlined in red in Figure 8) out of 83 enzymes defined in 544 unigenes probably encode enzymes involved in starch and sucrose metabolism (Figure 8). The processes comprising the starch and sucrose metabolic pathways included pentose and glucuronate interconversions, amino sugar and nucleotide sugar metabolism, ascorbate metabolism, retinol metabolism and glycolysis/gluconeogenesis. The other key enzymes in these pathways, their expression patterns and regulation during bulb development remain to be determined.

These transcript database and functional annotation acquired by RNA-seq provide a valuable resource for investigating specific processes, functions and pathways in *Lycoris*. For other plants, as far as we know, Zhang et al. (2010) presented the transcriptome atlas for eight organs of cultivated rice using high-throughput paired-end RNA-seq, with extremely-low-level-expressed transcripts and a substantial number of novel transcripts, exons and untranslated regions detected. The constitutive and alternative splicing was cataloged for *A. thaliana* using the Illumina RNA-seq approach (Filichkin et al., 2010), providing unparalleled evaluation of alternative splicing and confirming that more intron-containing than previous estimates were alternatively spliced. Besides, deep sequencing of transcriptome is also used for *A. annua* to identify genes and markers for certain breeding purpose (Graham et al., 2010). This transcriptome database of *L. sprengeri* will aid in functional genomic studies of *Lycoris* and facilitate further

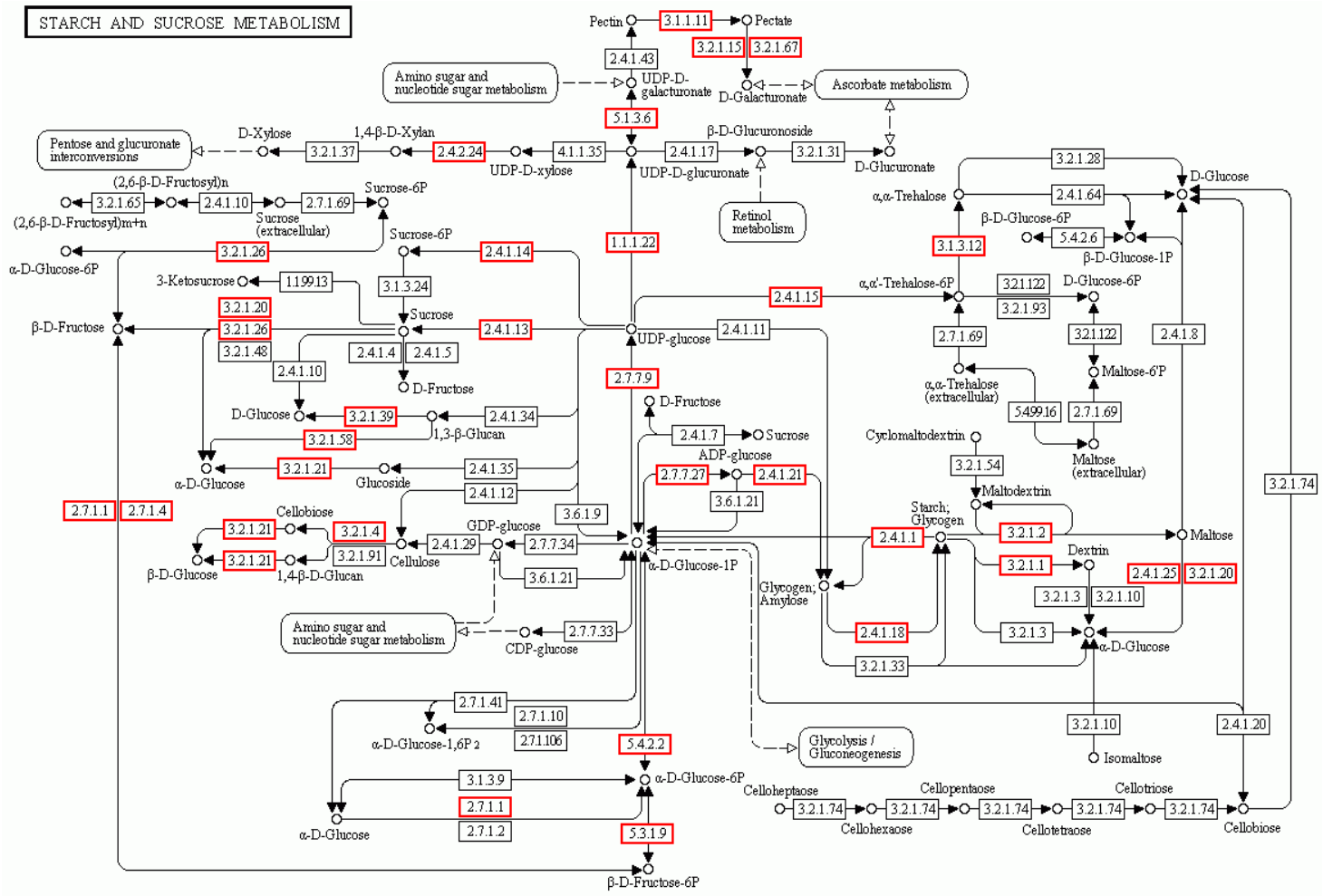

Figure 8. The starch and sucrose metabolic pathways.

molecular breeding methods.

REFERENCES

't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHA, Menezes RX, Boer JM, Ommen GJB, Dunnen JT (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. 36(21): 141.

Ansorge WJ (2009). Next-generation DNA sequencing techniques. New biotechnol. 25: 195-203.

Bryan J (1992). Manual of Bulbs. Timber Press, Portland Oregon.

Cirulli ET, Singh A, Shianna KV, Ge DL, Smith JP, Maia JM, Heinzen EL, Goedert JJ, Goldstein DB (2010). Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. Genome Biol. 11: 57.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21: 3674-3676.

Cui YL (2004). Construction of cDNA Library and Analysis on Expressed Sequence Tags of Genus of *Lycoris* Plant. DS Thesis, Nanjing Foresty University.

Delseny M, Han B, Hsing YL (2010). High throughput DNA sequencing: The new sequencing revolution. Plant Sci., 179: 407-422.

Filichkin SA, Priest HD, Givan SA, Shen RK, Bryant DW, Fox SE, Wong WK, Mockler TC (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res. 20: 45-58.

Graham IA, Besser K, Blumer S, Branigan CA, Czechowski T, Elias L, Guterman I, Harvey D, Isaac PG, Khan AM, Larson TR, Li Y, Pawson T, Penfield T, Rae AM, Rathbone DA, Reid S, Ross J, Smallwood MF, Segura V, Townsend T, Vyas D, Winzer T, Bowles D (2010). The genetic map of *Artemisia auuna* L. identifies loci affecting yield of the antimalarial drug Artemisinin. Science, 327: 328-331.

He QG, Kuang JF, Chen JY (2011). The role of lipoxygenase in banana fruit ripening. Scientia Agricultura Sinica. 44(1): 118-124.

He QL (2006). Construction of cDNA library and analysis on expressed sequence tags of genus of *Lycoris longituba*. MS Thesis, Nanjing Forestry University, Nanjing, China.

Rosenkranz R, Borodina T, Lehrach H, Himmelbauer H (2008). Characterizing the mouse ES cell transcriptome with Illumina sequencing. Genomics, 92: 187-194.

Schuster SC (2008). Next- generation sequencing transforms today's biology. Nature Methods, 5: 16-18.

Sun HM, Li TL, Li YF (2005). Changes of carbohydrate and amylase in lily bulb during bulb development. Bull. Botanical Res. 25: 59-63.

Wang L, Tang GG, Liu T (2009). Effects of 6-BA, GA_3 and NAA on flowering and endogenous hormone conent of *Lycoris sprengeri*. J. Nanjing Forestry University (Natural Sci. Edition), 33: 131-134.

Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS (2010). De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics, 11: 400.

Wang Z, Gerstein M, Snyder M (2009). RNA-seq: A reveolutionary tool for transcriptomics. Nat. Rev. Genet. 10: 57-63.
Xiang LX, He D, Dong WR, Zhang YW, Shao JZ (2010). Deep sequencing-based trancriptome profiling analysis of bacteria-challenged *Lateolabrax japonicus* reveals insight into the immune-relevant genes in marine fish. BMC Genomics, 11: 472.
Xue J, Bao YY, Li BL, Cheng YB, Peng ZY, Liu H, Xu HJ, Zhu ZR, Lou YG, Cheng JA, Zhang CX (2010). Transcriptome analysis of the brown planthopper *Nilaparvata lugens*. PLOS one, 5(12): 14233.
Ye J, Fang L, Zheng HK, Zhang Y, Chen J, Zhang ZJ, Wang J, Li ST, Li RQ, Bolund L, Wang J (2006). WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 34: 293-297.
Zhang GJ, Guo GW, Hu XD, Zhang Y, Li QY, Li RQ, Zhuang RH, Lu ZK, He ZQ, Fang XD, Chen L, Tian W, Tao Y, Kristiansen K, Zhang XQ, Li SG, Yang HM, Wang J, Wang J (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res. 20: 646-654.